CHARACTERIZATION OF THE HETEROGENEOUS PROGNOSTIC VALUE OF VARIOUS MUTATIONAL BURDEN ESTIMATES ACROSS THE TUMORS IN THE TCGA COHORT

by
Julia Kung

A thesis submitted to Johns Hopkins University in conformity with the requirements for the

degree of Master of Science

Baltimore, Maryland
August 2020

Abstract

Total mutation burden (TMB) is an aggregate genomic metric from somatic mutational data. The usage of TMB as a predictive biomarker of response to immunotherapy has been recently proposed but remains a contentious topic. Most of the previous works on TMB has focused on predicting checkpoint blockade response, but few have properly evaluated TMB as a prognostic factor. However, it is critical to understand any prognostic aspects of a biomarker before characterizing them as predictive biomarker to a specific therapeutic indication.

In this study, we analyzed curated clinical and survival outcome data from The Cancer Genome Atlas (TCGA) and corresponding somatic mutation data from the Multi-Center Mutation Calling in Multiple Cancers (MC3) project. We characterized a variety of mutation burden metrics on patient outcome using Inverse Probability Treatment Weighted (IPTW) Cox Proportional-Hazards model. We also combined the logic and analytics of the Cox Proportional-Hazards model with neural networks to explore if complex non-monotonic relationships between mutational burden estimated and patient outcome were present.

Our results show the associations between mutation burden and outcome, with both positive and negative effects of higher TMB observed depending on the tumor type. For

most tumor types, the results from our neural network models were congruent with the results from the conventional Cox modeling in that generally monotonic relationships were observed. However, we identified a few tumor types in which clear non-monotonic relationships existed, which could not be adequately characterized by conventional Cox modeling of these data.

In conclusion, the associations between mutation burden and outcome are tumor-specific, with both positive and negative effects of higher TMB observed depending on the tumor type. Understanding this background prognostic effect is critical in characterizing the utility of these metrics at predicting response to any given therapeutic intervention. Additionally, some tumor types exhibited non-monotonic relations between mutation burden metrics and outcome, stressing the importance of better understanding the nature of the prognostic information encoded in these mutation burden estimates before characterizing them in any treatment-predictive context.

Primary Reader and Advisor: Alexander Baras

Secondary Reader: Harold Lehmann

Acknowledgements

I am very grateful to my advisor, Dr. Alexander Baras, for introducing me to bioinformatics research, and for his overall guidance through my Master's study. Dr. Baras' profound knowledge and understanding of research methods have been a fantastic resource for me to learn from. I am also very thankful for him to always being patient with my learning process and guided me through numerous attempts and errors.

I am also grateful to my program director, Dr. Harold Lehmann, for letting me explore the wonderful world of Health Sciences Informatics. Dr. Lehmann always emphasizes the importance of selecting the correct method, and always encourages us to think deeper than the question on the surface. Under his guidance, the knowledge and skills I have learned in two years exceeded my expectation.

Last, I want to thank Yun-Fei Liu, a PhD student in Psychological & Brain Sciences in Johns Hopkins University. Yun-Fei helped me solve countless programming issues and also taught me how to troubleshoot. His help, support, and his brilliant ways of solving problems have been the motivation to me to improve my skills in analytical programming.

Table of Contents

List of Tables

List of Figures

Introduction

Total mutation burden (TMB) is an aggregate genomic metric from somatic mutational data. It is a quantitative measure of the total number of somatic mutations (usually qualified by some characteristic, such as non-synonymous mutations) found in a cancer cell's genome relative to the total number megabase pairs (Mbp) examined. This reflects the accumulation of somatic mutations over the tumor's lifetime, including both passenger along with driver mutations. Somatic mutations include but are not limited to in-frame insertion/ deletion, frameshift insertion/ deletion, missense mutation, nonsense mutation, and non-stop mutation. In most studies, the total number of coding mutations is normalized by the size of the exome (roughly 33 Mbps).

Other aggregate genomic metrics include biomarkers such as Microsatellite Instability (MSI) and other mutational signatures. MSI is characterized by changes in the lengths of representative elements in the genome at well-known loci; mutational signatures are generally defined by distinguishing mutation processes that involves unique components of DNA repair, replication, destruction, or modification.[1]

The usage of TMB as a predictive biomarker of response to immunotherapy has been recently proposed but remains a contentious topic. Studies have shown that higher TMB is

associated with better response to checkpoint inhibition treatment in several cancer types.[2-6] Additionally, patients with higher frameshift burden were also shown to have better checkpoint inhibitor response.[7] Moreover, the Food and Drug Administration (FDA) approved pembrolizumab treatment - a PD-1 inhibitor - for patients with high TMB in solid tumors. Higher TMB was defined in that study as greater than 10 TMB per Mbp as assessed from a gene panel assay only covering a small fraction of the exome.[8] Most of the previous works on TMB has focused on predicting checkpoint blockade response; but few have properly evaluated TMB as a prognostic factor. However, it is critical to understand any prognostic aspects of a biomarker before characterizing them as predictive biomarker to a specific therapeutic indication.

To characterize the prognostic effect of TMB, we performed survival analysis using clinical and molecular data obtained from The Cancer Genome Atlas (TCGA). TCGA contains 11,160 molecular profiles of tumor samples across 33 cancer types, with patient demographics and follow-up information for most cases.[9] Most of these cases represent early to middle stages of disease with effectively none of these cases having been treated with contemporary checkpoint blockade. As such the TCGA cohort is well suited to characterize the prognostic impact of TMB independent of its putative predictive capacity in the context of immunotherapy.

To assess the potential of TMB as a prognostic factor, we used Cox Proportional-Hazards model to evaluate the association between TMB and clinical endpoints. Cox Proportional Hazards-model is a regression model widely used in clinical research for investigating the relation between patient survival time and one or more covariates. It has been a valid cornerstone of survival-outcome correlative analysis. To enhance this model, we combined the flexibility of neural networks with Cox Proportional-Hazards model to identify more complex relationships (such as non-monotonic relationships) without requiring a priori transformations of the input predictors to these models.

Methods

Study Design

Curated and filtered clinical and survival outcome data were obtained from The Cancer Genome Atlas Pan (TCGA). TCGA contains molecular profiles of tumors from 11,160 patients across 33 cancer types, and includes demographics and four major clinical outcome endpoints: overall survival (OS), progression-free interval (PFI), disease-free interval (DFI), and disease-specific survival (DSS).[9]

Somatic mutation data that corresponds with the TCGA data were obtained from the Multi-Center Mutation Calling in Multiple Cancers (MC3) project. The MC3 project developed unified pipelines for variant calling and filtering methods to ensure quality and consistency. An extensive compilation of somatic mutation calls for the TCGA data were generated, which contains 10,295 tumor-normal pairs from 33 cancer types, with 3,600,963 produced somatic variants.[10]

We observed that some tumors with different histological characteristics were merged into one TCGA tumor type code. For example, Infiltrating Ductal Carcinoma and Infiltrating Lobular Carcinoma were both categorized into Breast Invasive Carcinoma (BRCA), while they do not share the same histological and clinical features.[11] For a closer reflection of tumor

biology, we re-labeled the tumor types using terminologies of National Cancer Institute Thesaurus (NCIt). [Appendix I: TCGA-NCIt corresponding table]

Of the 11,160 TCGA patient data, we excluded some records using the following criteria: (1) patient records that lacked information for NCIt re-labeling; (2) patient records that were not in the MC3 data; (3) patient records with incomplete clinical outcome endpoints (missing OS, OS time, PFI, or PFI time). Also, for the validity of outcome data analyses, we only included NCIt-labeled tumors with patient numbers higher than 100 and median follow-up time longer than 1 year (specifically 365 days). This criteria at the NCIt tumor type level filtered out the following (NCIt code in parentheses): Testicular Seminoma (C7328), Testicular Non-Seminomatous Germ Cell Tumor (C9313), Esophageal Adenocarcinoma (C4025), Pancreatic Carcinoma (C3850), Cervical Adenosquamous Carcinoma (C4519), Liposarcoma (C3194), Leiomyosarcoma (C3158), Mesothelioma (C3234), Cholangiocarcinoma (C4436), Breast Carcinoma (C4872), Endometrial Mixed Cell Adenocarcinoma (C40153), Thyroid Gland Follicular Carcinoma (C8054), Cervical Adenocarcinoma (C4029), Myxofibrosarcoma (C6496), Undifferentiated Pleomorphic Sarcoma (C4247), Synovial Sarcoma (C3400), Prostate Adenocarcinoma (C2919), Diffuse Large B-Cell Lymphoma (C8851), Esophageal Squamous Cell Carcinoma (C4024), Chromophobe Renal Cell Carcinoma (C4146), Uterine Carcinosarcoma (C42700), Adrenal Cortex Carcinoma (C9325), Paraganglioma (C3308),

Malignant Peripheral Nerve Sheath Tumor (C3798), Desmoid-Type Fibromatosis (C9182), and Uveal Melanoma (C7712).

After the selection criteria were applied, there were a total 8670 remained included patients across 25 NCIt tumor types, as listed below (NCIt code in parentheses): Adenocarcinoma, Pancreas (C8294), Astrocytoma (C60781), Cervical Squamous Cell Carcinoma (C4028), Clear Cell Renal Cell Carcinoma (C4033), Colorectal Adenocarcinoma (C5105), Cutaneous Melanoma (C3510), Endometrial Endometrioid Adenocarcinoma (C6287), Endometrial Serous Adenocarcinoma (C27838), Gastric Adenocarcinoma (C4004), Glioblastoma (C3058), Head and Neck Squamous Cell Carcinoma (C34447), Hepatocellular Carcinoma (C3099), Infiltrating Ductal Breast Carcinoma (C4194), Invasive Lobular Breast Carcinoma (C7950), Lung Adenocarcinoma (C3512), Lung Squamous Cell Carcinoma (C3493), Muscle-Invasive Bladder Carcinoma (C150572), Oligoastrocytoma (C4050), Oligodendroglioma (C3288), Ovarian Serous Adenocarcinoma (C7550), Papillary Renal Cell Carcinoma (C6975), Pheochromocytoma (C3326), Prostate Acinar Adenocarcinoma (C5596), Thymoma (C3411), and Thyroid Gland Papillary Carcinoma (C4035).

Definition

Using the MC3 somatic mutation data, we defined tumor mutation burden (TMB) as all identified variations; nonsynonymous (NonSyn) mutation as in-frame insertion/ deletion, frameshift insertion/ deletion, missense mutation, nonsense mutation, non-stop mutation, and RNA mutation; single nucleotide polymorphism (SNP) as missense mutation and nonsense mutation; and insertion-deletion (INDEL) mutation as in-frame insertion/ deletion and frameshift insertion/ deletion. All numbers of mutations were calculated per megabase pair (Mbp). Since all these data are derived from whole exome sequencing, these metrics were normalized by the size of the exome, which is approximately 33 Mbps.

From the TCGA data, we selected OS and PFI as clinical outcome endpoints. OS was defined as the duration from diagnosis date to date of death of any cause. It was a more accurate endpoint indicated by TCGA, as the events were unambiguously defined by the date of patient death. PFI was defined as the duration from the diagnosis date to date of the first occurrence of a new tumor event (includes disease progression, locoregional recurrence, distant metastasis, new primary tumor, or death with tumor). It was the suggested clinical endpoint choice for most of the cancer types in TCGA data, as PFI required relatively shorter clinical follow-up time, and more events were recorded within the study period.[9]

Statistical Analysis

To take account for the association between covariates and TMB, Propensity Score was produced from a generalized linear model. Following the two-step selection criteria of Wu et.al, potential covariates, including age, gender (male and female), and race (White, Black, Asian, and Others), were filtered for the model construction. For the univariate initial inclusion, a threshold of $p < 0.2$ was required. For the multivariate secondary inclusion, a threshold of $p < 0.1$ was required.[12] Using the generated Propensity Score, an Inverse Probability of Treatment Weight (IPTW) was evaluated for individual patients to compute the impact of TMB to OS and PFI in the context of a model that can account for covariate of TMB.

IPTW-weighted Cox Proportional-Hazards Model was then carried out to evaluate the impact of TMB to OS and PFI. Similar analyses were executed to assess the impact of NonSyn/ SNP/ INDEL to patient outcome (OS/ PFI).

By the result of Cox Proportional-Hazards model, we classified the tumor types into three groups based on the 90% confidence interval and the effect of increased mutations counts on the clinical endpoints (OS and PFI): positive effect (better outcome with increased mutation counts), negative effect (worse outcome with increased mutation counts), and no

significant association observed for tumor types that had confidence intervals crossing 1 (the line of no difference).

Neural Network

We used a simple 3 layer fully connected network that outputs a hazard score. The input layer into that final output layer of the network is interpreted as a vectors of predictors and the cox proportion hazard model is applied therein in which the loss function to be minimized is the negative log partial likelihood of the Cox model.[13] In this manner, which the model can characterize potential non-monotonic relationships between in input scalar measure, such as TMB, and survival measures.

We identified putative examples of such non-monotonic relationships in Endometrial Endometrioid Adenocarcinoma, Head and Neck Squamous Cell Carcinoma, and Lung Squamous Cell Carcinoma by examining the curvature of the neural network hazard output score across a range of input TMB.

For each tumor, we separate patients into three groups by the Hazard Ratio output of the Neural Network model, and verified the sufficiency of patient numbers in each group. The

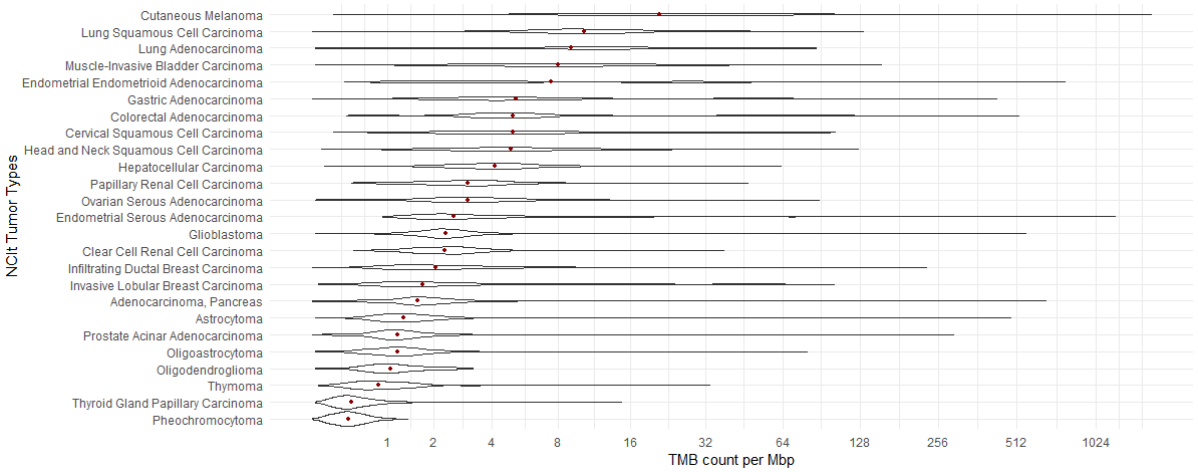
Kaplan Meier method was then performed, combined with p value computed by log-rank test to validate the non-monotonic relation between TMB and OS in actual patient data.

In this study, we carried out data manipulation and statistical analyses using R version 3.6.0. A p value < 0.05 was considered statistically significant.

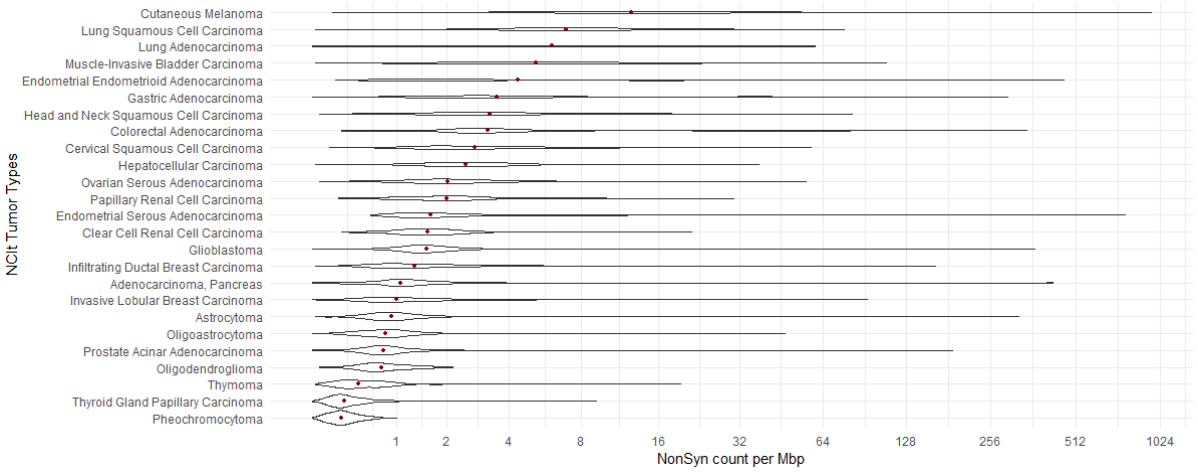
Results

After the tumor type selection criteria were applied, 8670 patients across 25 NCIt tumor types were included in this study. The patient demographics of included and excluded tumor types along with the aggregate mutation metrics (TMB/ NonSyn/ SNP/ INDEL) by each NCIt tumor type were shown in Appendix II and III. The median of TMB count ranged from 14 (0.42 per Mbp) of Pheochromocytoma to 692 (20.97 per Mbp) of Cutaneous Melanoma; the median of NonSyn count ranged from 9 (0.27 per Mbp) of Pheochromocytoma to 415 (12.58 per Mbp) of Cutaneous Melanoma; the median of SNP count ranged from 8 (0.24 per Mbp) of Pheochromocytoma to 401 (12.15 per Mbp) of Cutaneous Melanoma; last, the median of INDEL count ranged from 0 (0.00 per Mbp) of Pheochromocytoma and Thyroid Gland Papillary Carcinoma to 9 (0.27 per Mbp) of Endometrial Endometrioid Adenocarcinoma. (Figure 1A, 1B, 1C, 1D)

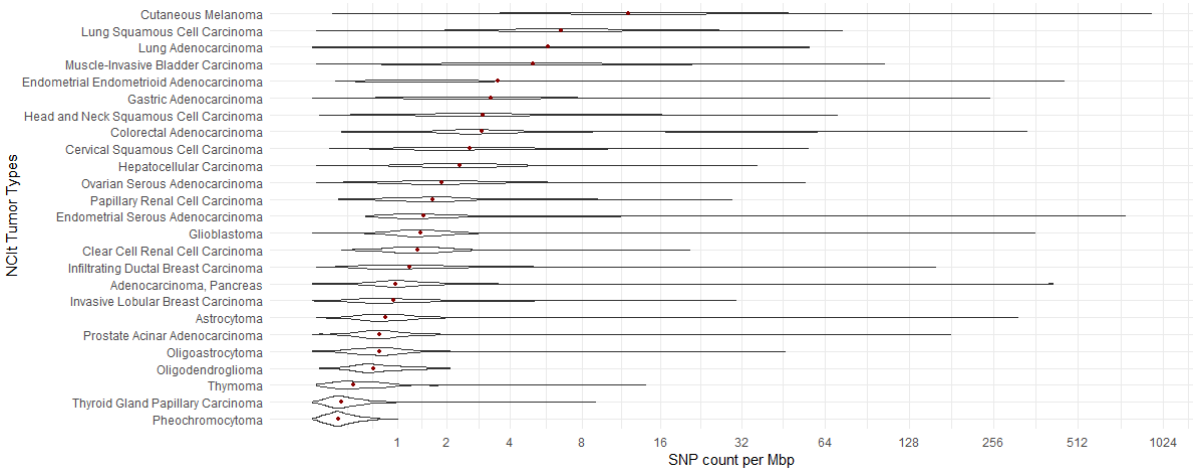
A. TMB



B. NonSyn



C. SNP



D. INDEL

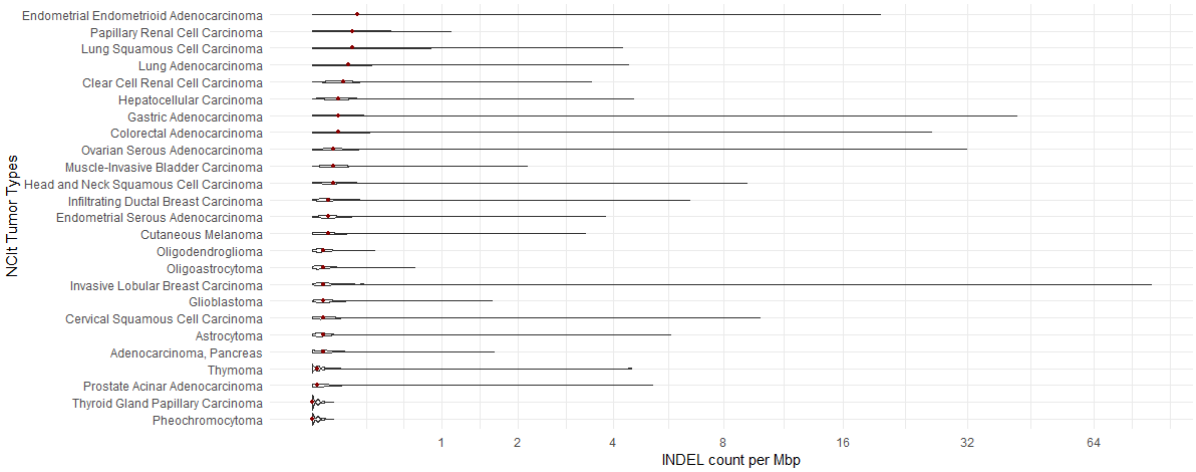
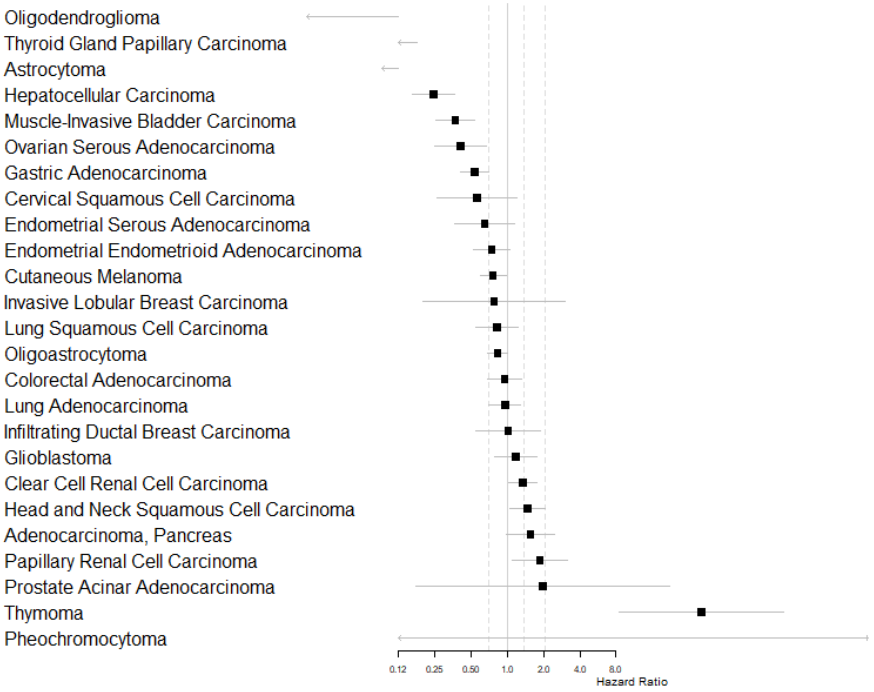


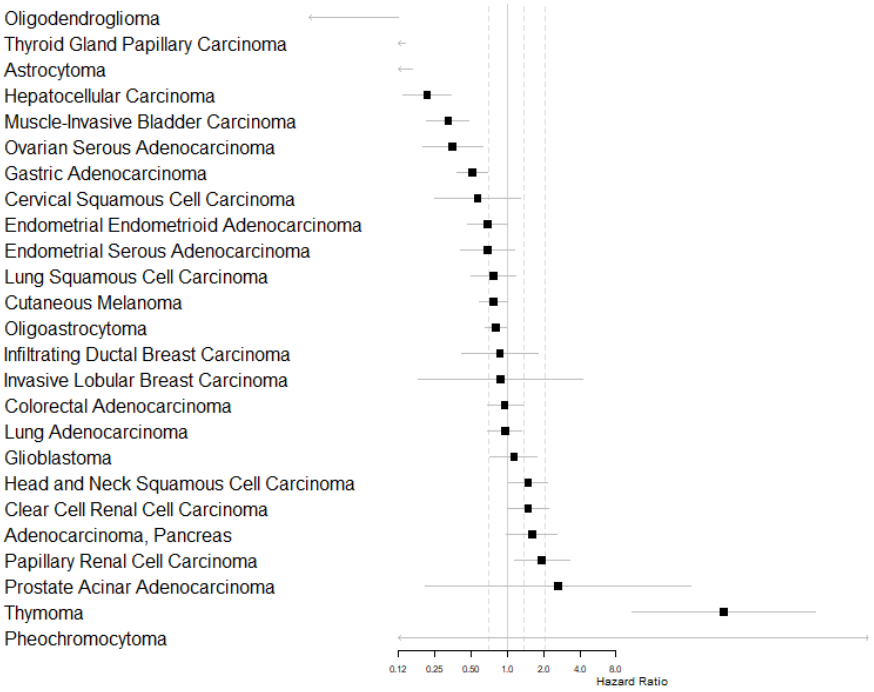
Figure 1 TMB, NonSyn, SNP, INDEL occurrence across 25 tumor types. The red dot in each violin plot indicates the median.

The impact of the mutation metrics (TMB/ NonSyn/ SNP/ INDEL) to OS and PFI were evaluated separately by IPTW-weighted Cox Proportional-Hazards model. The hazard ratios for these various genomic predictors across 25 tumor types are shown in Figure 2A, 2B, 2C, and 2D for OS as the clinical endpoint, and Figure 3A, 3B, 3C, and 3D for PFI as the clinical endpoint. We were able to classify the tumor types into three groups based on the 90% confidence interval (CI) and the effect of increased mutations counts on the clinical endpoints (OS and PFI): positive effect (better outcome with increased mutation counts), negative effect (worse outcome with increased mutation counts), and no significant association observed for tumor types that had confidence intervals crossing 1 (the line of no difference). These strata were identified for each of the mutation counts we examined: TMB, NonSyn, SNP, and INDEL.

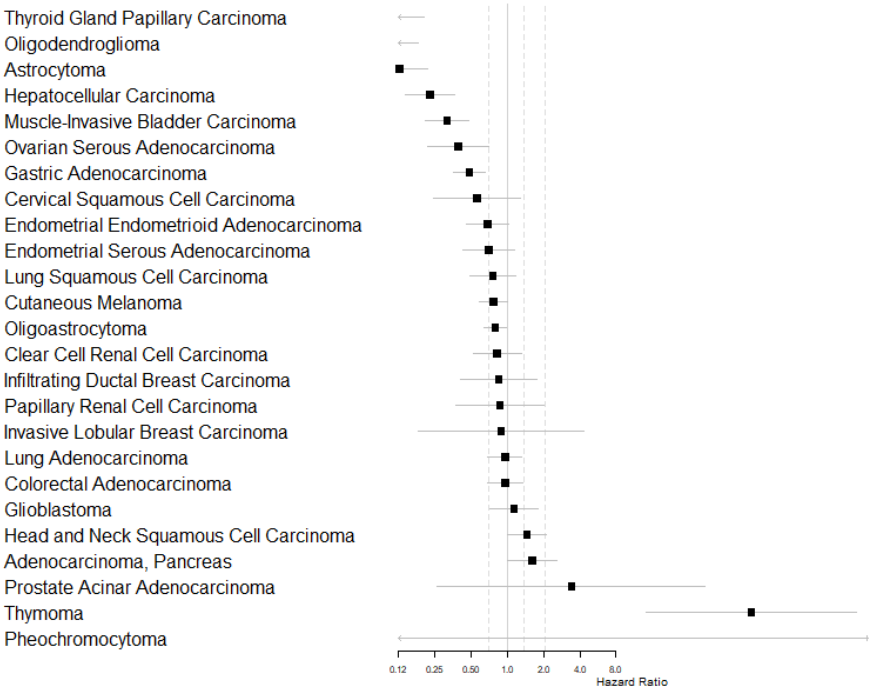
A. TMB



B. NonSyn



C. SNP



D. INDEL

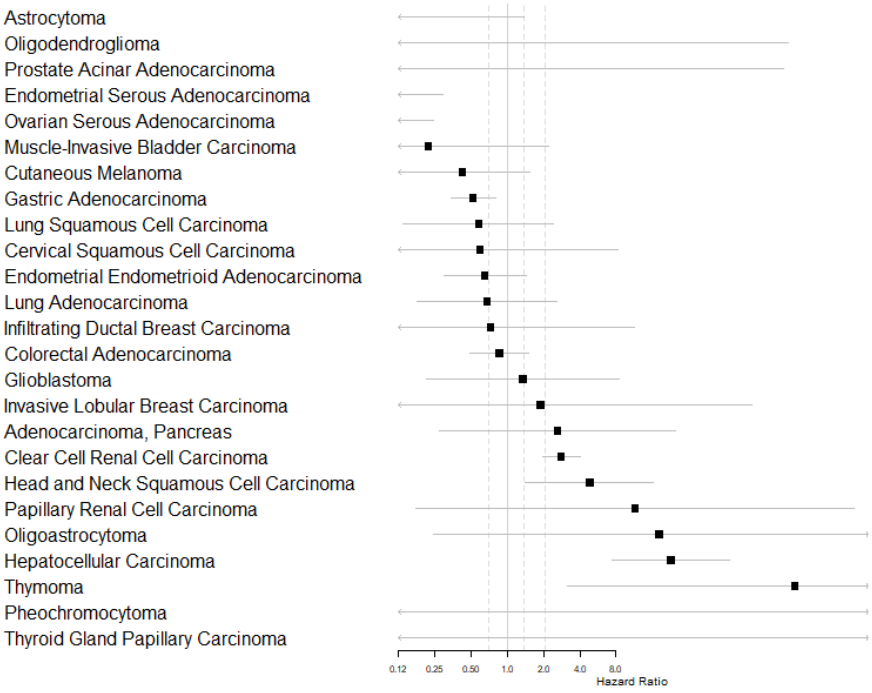
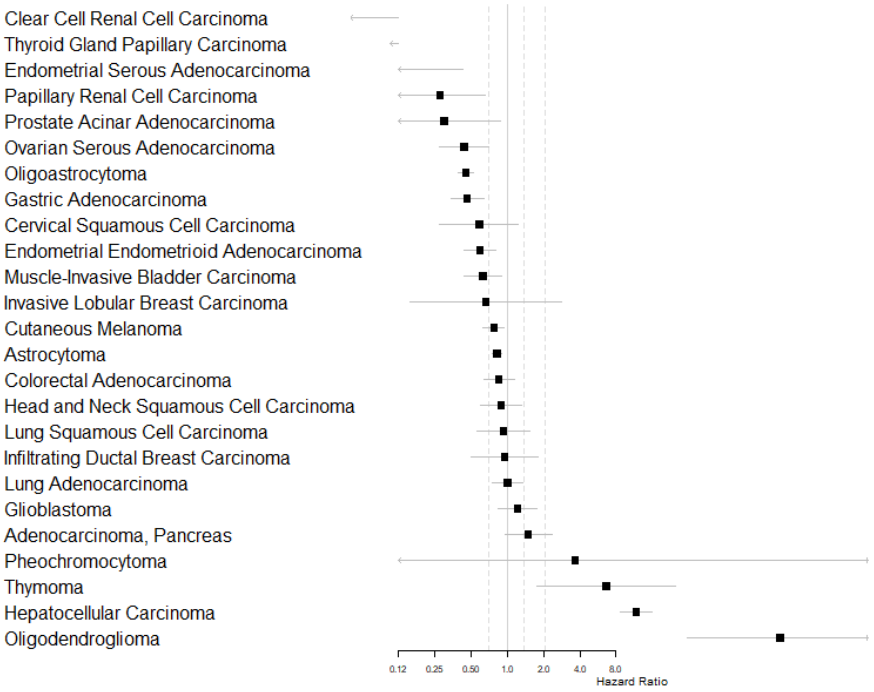
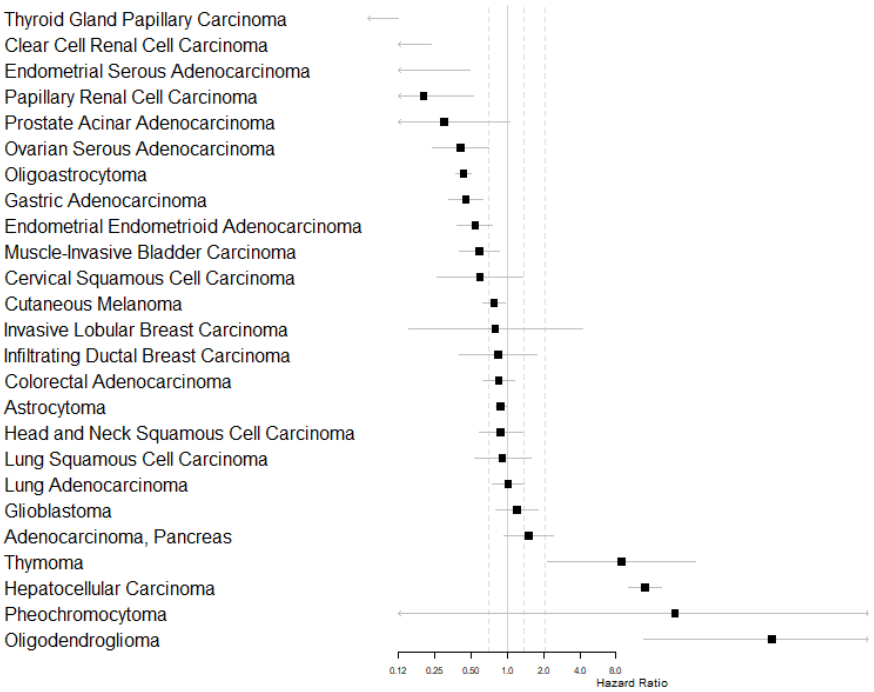


Figure 2. IPTW-weighted hazard ratios and 90% CI for the genomic predictors - TMB, NonSyn, SNP, INDEL - across 25 tumor types, using OS as the clinical endpoint.

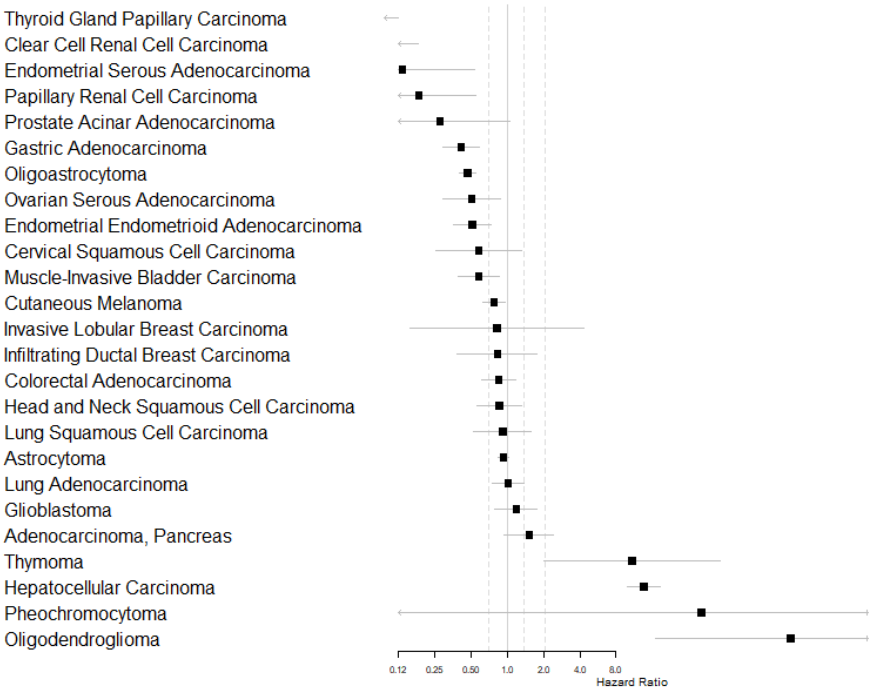
A. TMB



B. NonSyn



C. SNP



D. INDEL

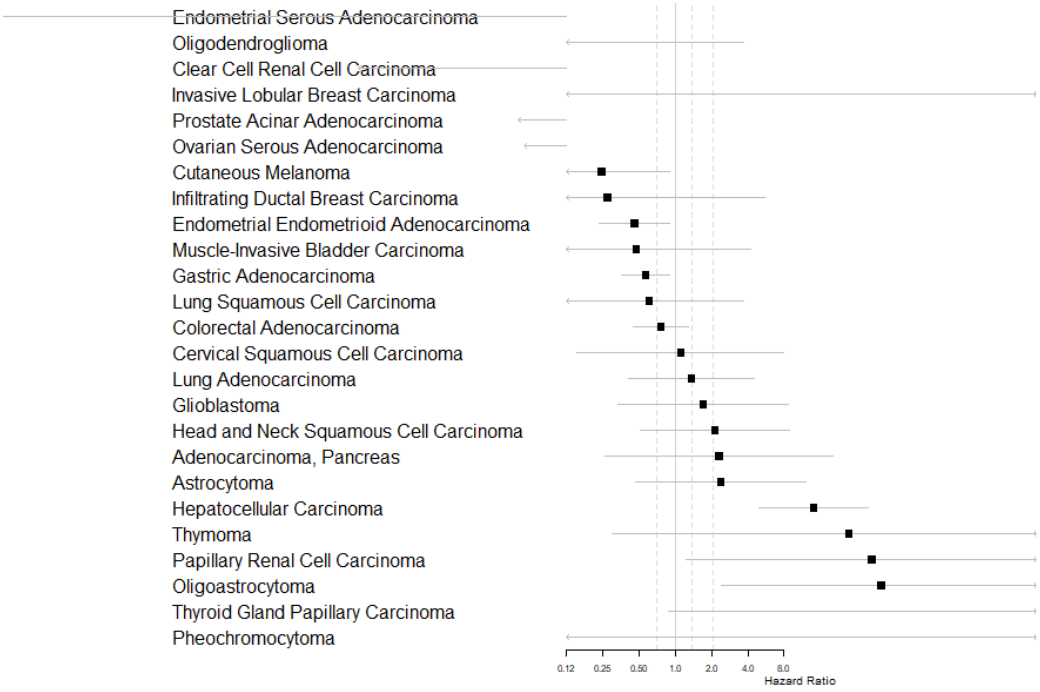


Figure 3. IPTW-weighted hazard ratios and 90% CI for the genomic predictors - TMB, NonSyn, SNP, INDEL - across 25 tumor types, using PFI as the clinical endpoint.

For the relation between TMB and OS (Fig 2A), the positive effect group included Oligodendroglioma, Thyroid Gland Papillary Carcinoma, Astrocytoma, Hepatocellular Carcinoma, Muscle-Invasive Bladder Carcinoma, Ovarian Serous Adenocarcinoma, Gastric Adenocarcinoma, and Cutaneous Melanoma. Oligodendroglioma had the lowest Hazard Ratio 0.004782 (95% CI 0.000787 – 0.029066).

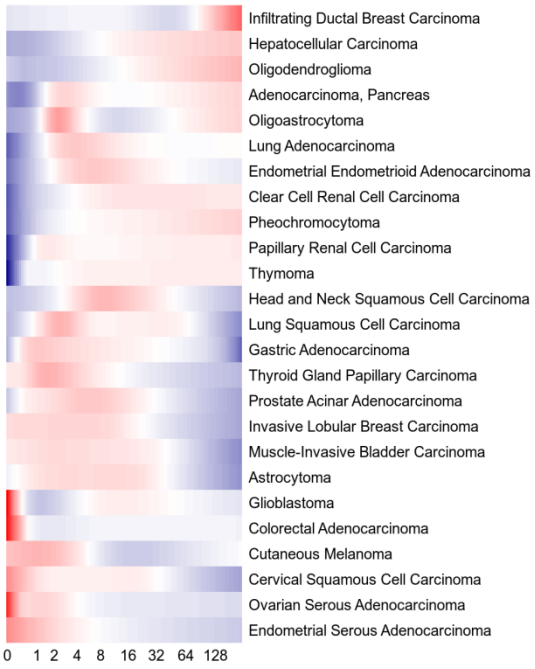
The negative effect group included Clear Cell Renal Cell Carcinoma, Head and Neck Squamous Cell Carcinoma, Papillary Renal Cell Carcinoma, and Thymoma. Thymoma had the highest Hazard Ratio 40.98 (95% CI 6.298 – 266.662).

The group of no significant association observed included Cervical Squamous Cell Carcinoma, Endometrial Serous Adenocarcinoma, Endometrial Endometrioid Adenocarcinoma, Invasive Lobular Breast Carcinoma, Lung Squamous Cell Carcinoma, Oligoastrocytoma, Colorectal Adenocarcinoma, Lung Adenocarcinoma, Infiltrating Ductal Breast Carcinoma, Glioblastoma, and Adenocarcinoma, Pancreas, and Prostate Acinar Adenocarcinoma.

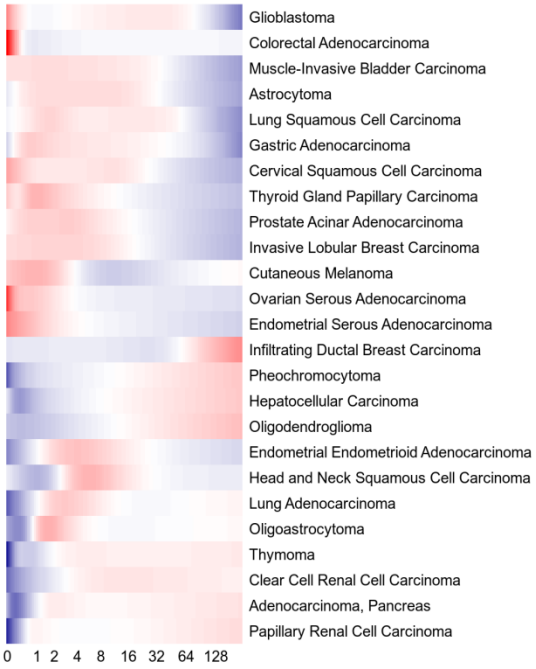
To further explore whether more complex relationships might exist between these mutation metrics and clinical endpoints, we performed the neural network model. The Hazard Ratio output of the model across 25 tumor types are shown in Figure 4A, 4B, 4C, 4D for OS as the clinical endpoint, and Figure 5A, 5B, 5C, 5D for PFI as the clinical endpoint. Most of the

tumors demonstrated monotonicity between the mutation counts and clinical outcome – either positive or negative effect - which is consistent with the outcome of the conventional Cox Proportional-Hazards model. However, some of the tumors exhibited non-monotonic relation between the mutation counts and clinical outcome. For example, of the impact on OS, Endometrial Endometrioid Adenocarcinoma, Head and Neck Squamous Cell Carcinoma, and Lung Squamous Cell Carcinoma displayed positive relation in lower TMB, negative relation in middle TMB, and positive relation again in higher TMB (Fig 4A).

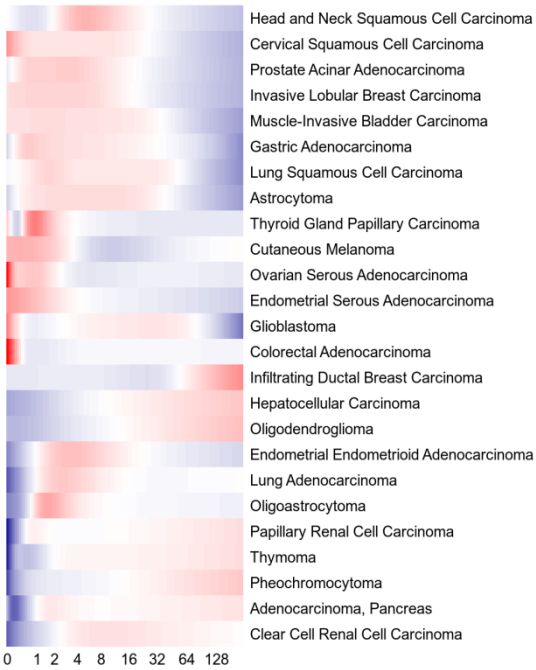
A. TMB



B. NonSyn



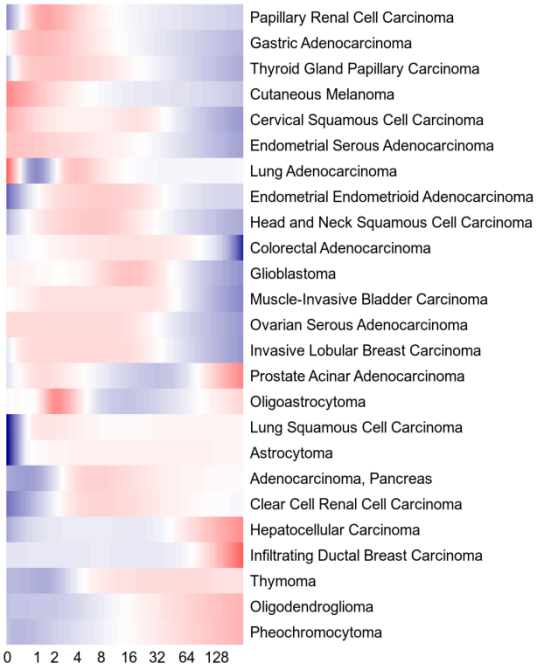
C. SNP



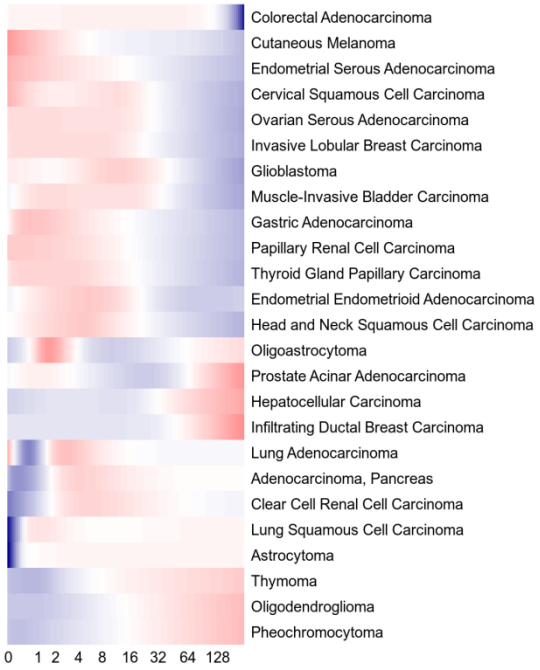
D. INDEL

Figure 4. Hazard Ratio output of the neural network model across 25 tumor types for OS as the clinical endpoint.

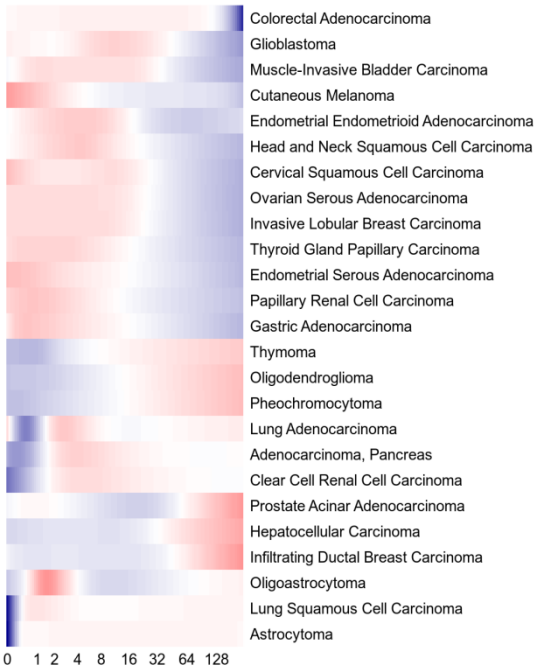
A. TMB



B. NonSyn



C. SNP



D. INDEL

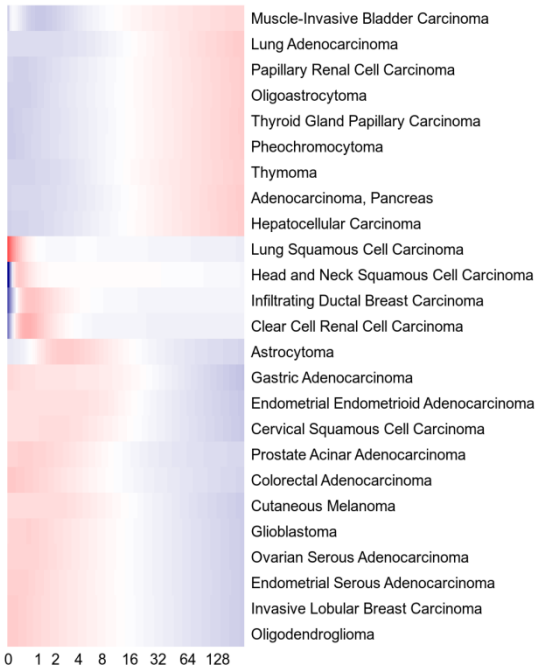
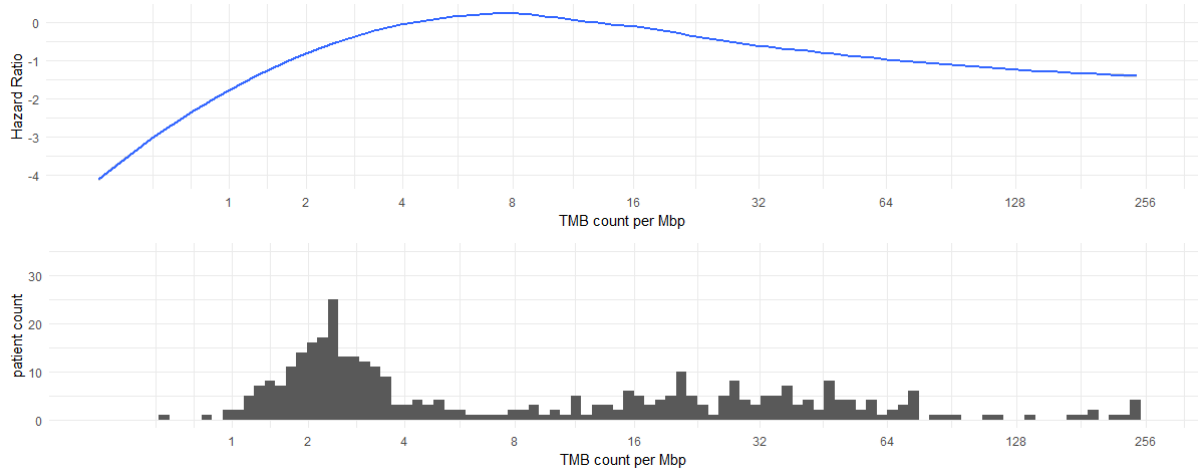


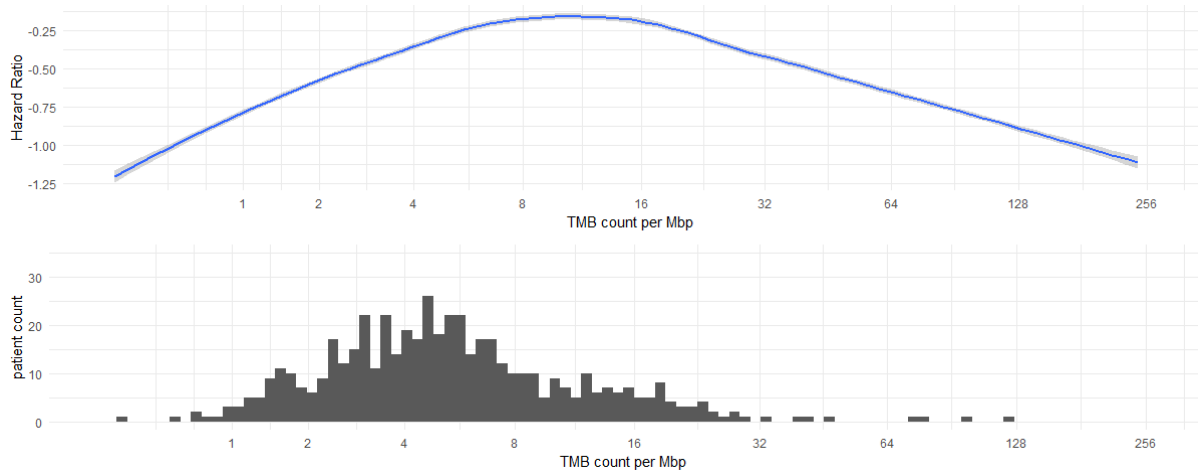
Figure 5. Hazard Ratio output of the neural network model across 25 tumor types for PFI as the clinical endpoint.

The identified tumors were then separately being examined and stratified into three groups by the predicted Hazard Ratios: low TMB, middle TMB, and high TMB. Among the three identified tumors with non-monotonic relation between TMB and OS, Lung Squamous Cell Carcinoma did not include sufficient patients in low and high TMB group (Fig 6C), while Endometrial Endometrioid Adenocarcinoma and Head and Neck Squamous Cell Carcinoma had valid number of patients in each group (Fig 6A, 6B). We then used the Kaplan Meier method accompanied with log-rank test for p value to analyze survival probability of the three groups. In Endometrial Endometrioid Adenocarcinoma, the middle TMB group has significantly lower (p value = 0.008) survival probability than low and high TMB group combined (Fig 7A). In Head and Neck Squamous Cell Carcinoma, the middle TMB group also demonstrated significantly lower (p value = 0.002) survival probability than low and high TMB group combined (Fig 7B).

A. Endometrial Endometrioid Adenocarcinoma



B. Head and Neck Squamous Cell Carcinoma



C. Lung Squamous Cell Carcinoma

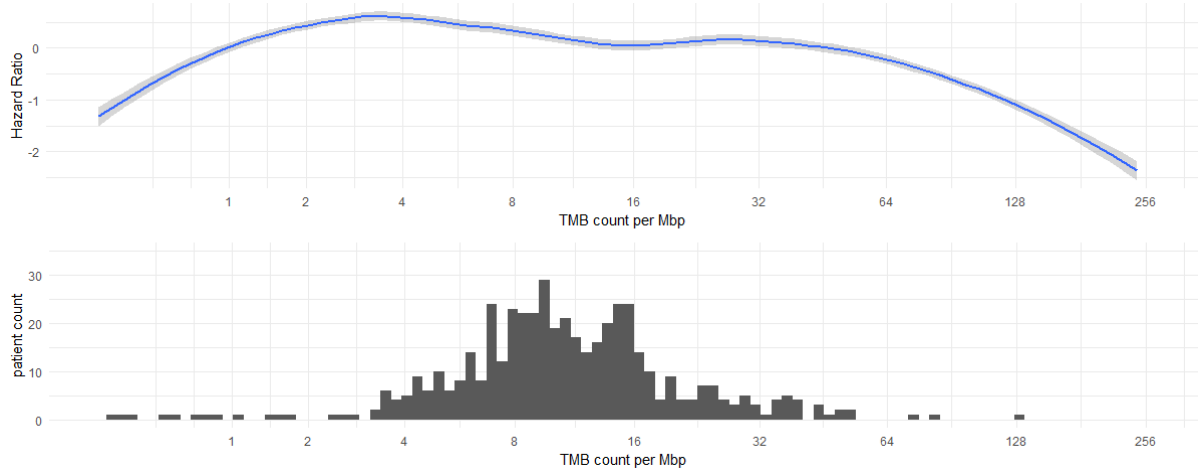
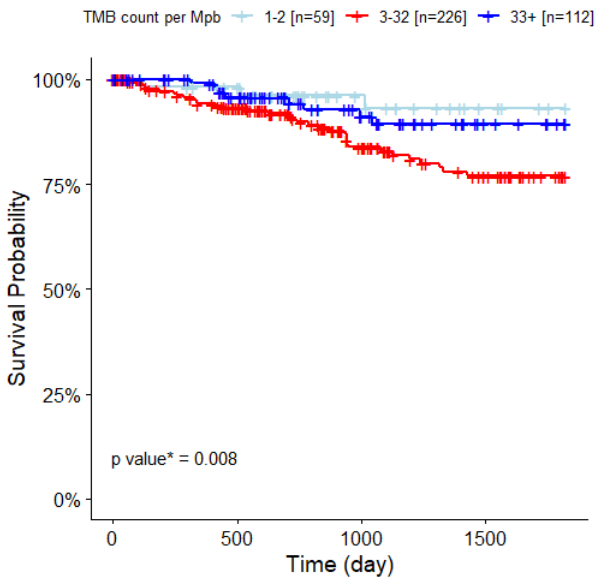


Figure 6. Hazard Ratio output (the impact of TMB to OS) by neural network model and patient distribution across TMB count per Mbp.

A. Endometrial Endometrioid Adenocarcinoma



B. Head and Neck Squamous Cell Carcinoma

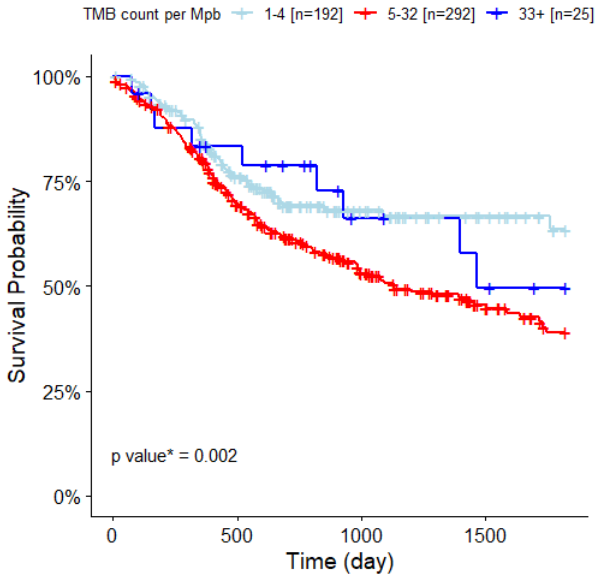


Figure 7. The Kaplan-Meier plots show the survival probability curves of low TMB, middle TMB, and high TMB groups in 5 years. *The p values of log-rank test showing combined low and high TMB versus middle TMB.

Discussion

In this study, we characterized a variety of mutation burden metrics on patient outcome. We analyzed curated clinical and survival outcome data from TCGA and corresponding somatic mutation data from the MC3 project. The TCGA tumor codes were re-labeled to better reflect clinically relevant groupings. For example, the histological subtypes of Breast Cancer - Infiltrating Ductal Carcinoma and Infiltrating Lobular Carcinoma - were both being categorized into Breast Invasive Carcinoma (BRCA) in TCGA data, while they differ in various histological, clinical, and radiographical features.[11,14] By leveraging the established ontology of the NCIt, we allow for more salient and clinically relevant conclusions to be drawn from the data.

A variety of mutation burden metrics and estimates have been proposed by some as biomarkers that can be used to predict the likelihood of response to checkpoint blockade immunotherapy across a variety of tumor types. However, we stress that without a baseline understanding of the underlying prognostic value of a biomarker, one cannot properly characterize its predictive nature. Our results from the IPTW-weighted Cox Proportional-Hazards Model highlight not only is it clear that there exists underlying prognostic information from these mutation burden metrics, but also that these effects are different across tumor types. In some tumor types, higher TMB counts were associated with

better outcomes (such as in Muscle-Invasive Bladder Carcinoma) while in other tumors higher TMB counts were associated with worse outcomes (such as in Clear Cell Renal Cell Carcinoma). The importance of these baseline characterization of prognostication in the context of a cohort (TCGA) not treated with checkpoint blockage immunotherapy cannot be over stressed. Consider, if the predictive effect of TMB in the context of an immunotherapy treated cohort is the same as what we observed in the TCGA, then it is very likely that in that context the biomarker is in fact not predictive of response to immunotherapy but rather represents a background prognostic information. Conversely, consider if the predictive effect of TMB in the context of immunotherapy were observed to be non-significant and its prognostic value in a non-immunotherapy treated cohort was shown to be correlated with worse outcome. One might false assume that it is not a biomarker of response/ benefit from immunotherapy, when it facts the opposite is true.

Having attempted to characterize potential prognostic information encoded in these mutation burden estimates using conventional biostatistics tools, we then sought to explore the possibility of more complex relationships between mutation burden metrics and patient outcomes. We leveraged the flexibility of neural networks to be able to learning complex encoding of data as they have been in medical diagnosis, sequence analysis, and cancer drug mechanism prediction.[15,16] We combined the logic and analytics of the Cox

Proportional-Hazards model with neural networks to explore if complex non-monotonic relationships between mutational burden estimated and patient outcome were present. For most tumor types, the results from our neural network model were congruent with the results from the conventional Cox modeling in that generally monotonic relationships were observed. However, intriguingly we identified a few tumor types in which clear non-monotonic relationships existed which could not be adequately characterized by conventional Cox modeling of these data. As was shown for Endometrial Endometroid Adenocarcinoma, patients whose tumors had either higher or lower TMB exhibited better outcomes, while intermediate levels of TMB correlated with worse outcomes. This may supported to some degree by the fact that two competing processes are in play in the biology here: (a) increased mutation rate may increase the likelihood of the tumor developing the ability to cause more severe disease (b) increased mutation rate may increase the likelihood of either altering a critical cellular system or generating a mutant form of a protein that would elicit an immune response.

One limitation in our study is that we did not directly use the Inverse Probability Treatment Weight (IPTW) in the Neural Network model. But, rather just wanted to explore whether more complex relationships (non-monotonic) might exist between these metrics and outcome. We highlighted the findings in Endometrial Endometrioid Adenocarcinoma and

Head and Neck Squamous Cell Carcinoma; however, better characterization and validation are required in our future work. Additionally, this work is based on the TCGA whole exome data while the vast majority of clinical testing is done using focus gene panels that only cover 5-10% of the exome. More work will be required to ascertain how well these findings would translate to conventional gene panel-based testing.

In conclusion, the associations between mutation burden and outcome are tumor-specific, with both positive and negative effects of higher TMB observed depending on the tumor type. Understanding this background prognostic effect is critical in characterizing the utility of these metrics at predicting response to any given therapeutic intervention. Additionally, some tumor types exhibited non-monotonic relations between mutation burden metrics and outcome, stressing the importance of better understanding the nature of the prognostic information encoded in these mutation burden estimates before characterizing them in any treatment-predictive context.

References

1. Alexandrov LB, Kim J, Haradhvala NJ, et al. The repertoire of mutational signatures in human cancer. *Nature (London)*. 2020;578(7793):94-101.

2. Havel JJ, Chowell D, Chan TA. The evolving landscape of biomarkers for checkpoint inhibitor immunotherapy. *Nature reviews. Cancer*. 2019;19(3):133-150.

3. Samstein RM, Lee C, Shoushtari AN, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature Genetics*. 2019;51(2):202-206.

4. Chalmers ZR, Connelly CF, Fabrizio D, et al. Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden. *Genome medicine*. 2017;9(1):34.

5. Rizvi NA, Hellmann MD, Snyder A, et al. Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. *Science (American Association for the Advancement of Science)*. 2015;348(6230):124-128.

6. Snyder A, Makarov V, Merghoub T, et al. Genetic basis for clinical response to CTLA-4 blockade in melanoma. *The New England journal of medicine*. 2014;371(23):2189-2199.

7. Turajlic S, Litchfield K, Xu H, et al. Insertion-and-deletion-derived tumour-specific neoantigens and the immunogenic phenotype: A pan-cancer analysis. *The lancet oncology*. 2017;18(8):1009-1021.

8. FDA approves pembrolizumab for adults and children with TMB-H solid tumors. U.S. Food & Drug Administration Web site. https://www.fda.gov/drugs/drug-approvals-and-databases/fda-approves-pembrolizumab-adults-and-children-tmb-h-solid-tumors. Updated 2020. Accessed Aug 21, 2020.

9. Liu J, Lichtenberg T, Hoadley KA, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*. 2018;173(2):400-416.e11.

10. Ellrott K, Bailey MH, Covington KR, et al. Scalable open science approach for mutation calling of tumor exomes using multiple genomic pipelines. *Cell systems*. 2018;6(3):271-281.e7.

11. Du T, Zhu L, Levine KM, et al. Invasive lobular and ductal breast carcinoma differ in immune response, protein translation efficiency and metabolism. *Sci Rep*. 2018;8(1):7205.

12. Wu H, Wang Z, Zhao Q, et al. Tumor mutational and indel burden: A systematic pan-cancer evaluation as prognostic biomarkers. *Annals of translational medicine*. 2019;7(22):640.

13. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: Personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*. 2018;18(1):24.

14. Thomas M, Kelly ED, Abraham J, Kruse M. Invasive lobular breast cancer: A review of pathogenesis, diagnosis, management, and future directions of early stage disease. *Seminars in Oncology*. 2019;46(2):121-132.

15. Faraggi D, Simon R. A neural network model for survival data. *Statistics in medicine*. 1995;14(1):73-82.

16. Wang P, Li Y, Reddy C. Machine learning for survival analysis. *ACM Computing Surveys (CSUR)*. 2019;51(6):1-36.

Appendices

Appendix 1: TCGA-NCIt Corresponding Table

TCGA tumor type	TCGA histological type	NCI-T Label	NCI-T Code	counts
ACC	Adrenocortical Carcinoma- Myxoid Type	Adrenal Cortex Carcinoma	C9325	1
ACC	Adrenocortical Carcinoma- Oncocytic Type	Adrenal Cortex Carcinoma	C9325	4
ACC	Adrenocortical carcinoma- Usual Type	Adrenal Cortex Carcinoma	C9325	87
BLCA	Muscle invasive urothelial carcinoma (pT2 or above)	Muscle-Invasive Bladder Carcinoma	C150572	409
BLCA	[Not Available]			3
BRCA	Infiltrating Carcinoma NOS	Breast Carcinoma	C4872	1
BRCA	Infiltrating Ductal Carcinoma	Infiltrating Ductal Breast Carcinoma	C4194	784
BRCA	Infiltrating Lobular Carcinoma	Invasive Lobular Breast Carcinoma	C7950	203
BRCA	Medullary Carcinoma	Breast Carcinoma	C4872	6
BRCA	Metaplastic Carcinoma	Breast Carcinoma	C4872	9
BRCA	Mixed Histology (please specify)	Breast Carcinoma	C4872	30
BRCA	Mucinous Carcinoma	Breast Carcinoma	C4872	17
BRCA	Other, specify	Breast Carcinoma	C4872	46
BRCA	[Not Available]			1
CESC	Adenosquamous	Cervical Adenosquamous Carcinoma	C4519	6
CESC	Cervical Squamous Cell Carcinoma	Cervical Squamous Cell Carcinoma	C4028	254
CESC	Endocervical Adenocarcinoma of the Usual Type	Cervical Adenocarcinoma	C4029	6
CESC	Endocervical Type of Adenocarcinoma	Cervical Adenocarcinoma	C4029	21

CESC	Endometrioid Adenocarcinoma of Endocervix	Cervical Adenocarcinoma	C4029	3
CESC	Mucinous Adenocarcinoma of Endocervical Type	Cervical Adenocarcinoma	C4029	17
CHOL	Cholangiocarcinoma; distal	Cholangiocarcinoma	C4436	2
CHOL	Cholangiocarcinoma; hilar/perihilar	Cholangiocarcinoma	C4436	7
CHOL	Cholangiocarcinoma; intrahepatic	Cholangiocarcinoma	C4436	36
COAD	Colon Adenocarcinoma	Colorectal Adenocarcinoma	C5105	392
COAD	Colon Mucinous Adenocarcinoma	Colorectal Adenocarcinoma	C5105	62
COAD	[Discrepancy]			3
COAD	[Not Available]			2
DLBC	Diffuse large B-cell lymphoma (DLBCL) NOS (any anatomic site nodal or extranodal)	Diffuse Large B-Cell Lymphoma	C8851	41
DLBC	Primary DLBCL of the CNS	Diffuse Large B-Cell Lymphoma	C8851	3
DLBC	Primary mediastinal (thymic) DLBCL	Diffuse Large B-Cell Lymphoma	C8851	4
ESCA	Esophagus Adenocarcinoma, NOS	Esophageal Adenocarcinoma	C4025	89
ESCA	Esophagus Squamous Cell Carcinoma	Esophageal Squamous Cell Carcinoma	C4024	96
GBM	Glioblastoma Multiforme (GBM)	Glioblastoma	C3058	31
GBM	Treated primary GBM	Glioblastoma	C3058	20
GBM	Untreated primary (de novo) GBM	Glioblastoma	C3058	545
HNSC	Head & Neck Squamous Cell Carcinoma	Head and Neck Squamous Cell Carcinoma	C34447	517
HNSC	Head & Neck Squamous Cell Carcinoma Basaloid Type	Head and Neck Squamous Cell Carcinoma	C34447	10
HNSC	Head & Neck Squamous Cell Carcinoma, Spindle Cell Variant	Head and Neck Squamous Cell Carcinoma	C34447	1
KICH	Kidney Chromophobe	Chromophobe Renal Cell Carcinoma	C4146	113
KIRC	Kidney Clear Cell Renal Carcinoma	Clear Cell Renal Cell Carcinoma	C4033	537
KIRP	Kidney Papillary Renal Cell Carcinoma	Papillary Renal Cell Carcinoma	C6975	291

LAML		Acute Myeloid Leukemia	C3171	200
LGG	Astrocytoma	Astrocytoma	C60781	194
LGG	Oligoastrocytoma	Oligodendroglioma	C3288	130
LGG	Oligodendroglioma	Oligoastrocytoma	C4050	191
LIHC	Fibrolamellar Carcinoma	Hepatocellular Carcinoma	C3099	3
LIHC	Hepatocellular Carcinoma	Hepatocellular Carcinoma	C3099	367
LIHC	Hepatocholangiocarcinoma (Mixed)	Hepatocellular Carcinoma	C3099	7
LUAD	Lung Adenocarcinoma	Lung Adenocarcinoma	C3512	522
LUSC	Lung Squamous Cell Carcinoma	Lung Squamous Cell Carcinoma	C3493	504
MESO	Biphasic mesothelioma	Mesothelioma	C3234	23
MESO	Diffuse malignant mesothelioma - NOS	Mesothelioma	C3234	5
MESO	Epithelioid mesothelioma	Mesothelioma	C3234	57
MESO	Sarcomatoid mesothelioma	Mesothelioma	C3234	2
OV	Serous Cystadenocarcinoma	Ovarian Serous Adenocarcinoma	C7550	587
PAAD	Pancreas-Adenocarcinoma Ductal Type	Adenocarcinoma, Pancreas	C8294	154
PAAD	Pancreas-Adenocarcinoma-Other Subtype	Adenocarcinoma, Pancreas	C8294	25
PAAD	Pancreas-Colloid (mucinous non-cystic) Carcinoma	Pancreatic Carcinoma	C3850	4
PAAD	Pancreas-Undifferentiated Carcinoma	Pancreatic Carcinoma	C3850	1
PAAD	[Discrepancy]			1
PCPG	Paraganglioma	Paraganglioma	C3308	18
PCPG	Paraganglioma; Extra-adrenal Pheochromocytoma	Paraganglioma	C3308	13
PCPG	Pheochromocytoma	Pheochromocytoma	C3326	148
PRAD	Prostate Adenocarcinoma Acinar Type	Prostate Acinar Adenocarcinoma	C5596	485
PRAD	Prostate Adenocarcinoma, Other Subtype	Prostate Adenocarcinoma	C2919	15

READ	Rectal Adenocarcinoma	Colorectal Adenocarcinoma	C5105	151
READ	Rectal Mucinous Adenocarcinoma	Colorectal Adenocarcinoma	C5105	13
READ	[Not Available]			6
SARC	Dedifferentiated liposarcoma	Liposarcoma	C3194	59
SARC	Desmoid Tumor	Desmoid-Type Fibromatosis	C9182	2
SARC	Giant cell 'MFH' / Undifferentiated pleomorphic sarcoma with giant cells	Undifferentiated Pleomorphic Sarcoma	C4247	1
SARC	Leiomyosarcoma (LMS)	Leiomyosarcoma	C3158	105
SARC	Malignant Peripheral Nerve Sheath Tumors (MPNST)	Malignant Peripheral Nerve Sheath Tumor	C3798	9
SARC	Myxofibrosarcoma	Myxofibrosarcoma	C6496	25
SARC	Pleomorphic 'MFH' / Undifferentiated pleomorphic sarcoma	Undifferentiated Pleomorphic Sarcoma	C4247	29
SARC	Sarcoma; synovial; poorly differentiated	Synovial Sarcoma	C3400	2
SARC	Synovial Sarcoma - Biphasic	Synovial Sarcoma	C3400	2
SARC	Synovial Sarcoma - Monophasic	Synovial Sarcoma	C3400	6
SARC	Undifferentiated Pleomorphic Sarcoma (UPS)	Undifferentiated Pleomorphic Sarcoma	C4247	21
SKCM		Cutaneous Melanoma	C3510	470
STAD	Stomach Adenocarcinoma, Signet Ring Type	Gastric Adenocarcinoma	C4004	13
STAD	Stomach, Adenocarcinoma, Diffuse Type	Gastric Adenocarcinoma	C4004	72
STAD	Stomach, Adenocarcinoma, Not Otherwise Specified (NOS)	Gastric Adenocarcinoma	C4004	164
STAD	Stomach, Intestinal Adenocarcinoma, Mucinous Type	Gastric Adenocarcinoma	C4004	22
STAD	Stomach, Intestinal Adenocarcinoma, Not Otherwise Specified (NOS)	Gastric Adenocarcinoma	C4004	82
STAD	Stomach, Intestinal Adenocarcinoma, Papillary Type	Gastric Adenocarcinoma	C4004	8

STAD	Stomach, Intestinal Adenocarcinoma, Tubular Type	Gastric Adenocarcinoma	C4004	79
STAD	[Discrepancy]			1
STAD	[Not Available]			2
TGCT	Non-Seminoma; Choriocarcinoma Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Teratoma (Mature) Non-Seminoma; Teratoma (Immature)	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Embryonal Carcinoma	Testicular Non-Seminomatous Germ Cell Tumor	C9313	15
TGCT	Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Teratoma (Mature)	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Teratoma (Mature) Non-Seminoma; Choriocarcinoma Non-Seminoma; Yolk Sac Tumor Seminoma; NOS	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Teratoma (Mature) Non-Seminoma; Teratoma (Immature) Non-Seminoma; Yolk Sac Tumor	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Yolk Sac Tumor	Testicular Non-Seminomatous Germ Cell Tumor	C9313	6
TGCT	Non-Seminoma; Embryonal Carcinoma	Testicular Non-Seminomatous Germ Cell	C9313	1

	Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Choriocarcinoma	Tumor		
TGCT	Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Teratoma (Immature)	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Teratoma (Mature)	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Teratoma (Mature) Non-Seminoma; Teratoma (Immature) Seminoma; NOS	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Embryonal Carcinoma Seminoma; NOS	Testicular Non-Seminomatous Germ Cell Tumor	C9313	2
TGCT	Non-Seminoma; Embryonal Carcinoma [Not Available]	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Embryonal Carcinoma [Not Available] [Not Available] [Not Available]	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Teratoma (Immature)	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Teratoma (Immature)	Testicular Non-Seminomatous Germ Cell	C9313	1

	Non-Seminoma; Teratoma (Mature) Non-Seminoma; Yolk Sac Tumor	Tumor		
TGCT	Non-Seminoma; Teratoma (Immature) Non-Seminoma; Yolk Sac Tumor	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Teratoma (Immature) Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Embryonal Carcinoma	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Teratoma (Immature) Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Teratoma (Mature)	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Teratoma (Mature)	Testicular Non-Seminomatous Germ Cell Tumor	C9313	2
TGCT	Non-Seminoma; Teratoma (Mature) Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Teratoma (Immature) Non-Seminoma; Yolk Sac Tumor	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Teratoma (Mature) Non-Seminoma; Teratoma (Immature) Non-Seminoma; Embryonal Carcinoma	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Teratoma (Mature) Non-Seminoma; Teratoma (Immature) Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Yolk Sac Tumor	Testicular Non-Seminomatous Germ Cell Tumor	C9313	4

TGCT	Non-Seminoma; Teratoma (Mature) Non-Seminoma; Teratoma (Immature) Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Choriocarcinoma	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Teratoma (Mature) Non-Seminoma; Teratoma (Immature) Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Embryonal Carcinoma Seminoma; NOS	Testicular Non-Seminomatous Germ Cell Tumor	C9313	2
TGCT	Non-Seminoma; Teratoma (Mature) Non-Seminoma; Yolk Sac Tumor	Testicular Non-Seminomatous Germ Cell Tumor	C9313	2
TGCT	Non-Seminoma; Teratoma (Mature) Seminoma; NOS Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Yolk Sac Tumor	Testicular Non-Seminomatous Germ Cell Tumor	C9313	2
TGCT	Non-Seminoma; Teratoma (Mature) [Not Available] Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Yolk Sac Tumor	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Yolk Sac Tumor	Testicular Non-Seminomatous Germ Cell Tumor	C9313	3
TGCT	Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Embryonal Carcinoma	Testicular Non-Seminomatous Germ Cell Tumor	C9313	2

TGCT	Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Teratoma (Mature) Non-Seminoma; Teratoma (Immature)	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Teratoma (Mature) Seminoma; NOS	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Teratoma (Mature)	Testicular Non-Seminomatous Germ Cell Tumor	C9313	1
TGCT	Seminoma; NOS	Testicular Seminoma	C7328	64
TGCT	Seminoma; NOS Non-Seminoma; Choriocarcinoma	Testicular Seminoma	C7328	1
TGCT	Seminoma; NOS Non-Seminoma; Embryonal Carcinoma	Testicular Seminoma	C7328	1
TGCT	Seminoma; NOS Non-Seminoma; Teratoma (Mature)	Testicular Seminoma	C7328	1
TGCT	Seminoma; NOS Non-Seminoma; Teratoma (Mature) Non-Seminoma; Embryonal Carcinoma	Testicular Seminoma	C7328	1
TGCT	Seminoma; NOS Non-Seminoma; Teratoma (Mature) Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Yolk Sac Tumor	Testicular Seminoma	C7328	1

TGCT	Seminoma; NOS Non-Seminoma; Teratoma (Mature) Non-Seminoma; Teratoma (Immature) Non-Seminoma; Yolk Sac Tumor	Testicular Seminoma	C7328	1
TGCT	Seminoma; NOS Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Embryonal Carcinoma Non-Seminoma; Teratoma (Immature)	Testicular Seminoma	C7328	1
TGCT	Seminoma; NOS Non-Seminoma; Yolk Sac Tumor Non-Seminoma; Teratoma (Mature) Non-Seminoma; Teratoma (Immature)	Testicular Seminoma	C7328	1
THCA	Other, specify			9
THCA	Thyroid Papillary Carcinoma - Classical/usual	Thyroid Gland Papillary Carcinoma	C4035	359
THCA	Thyroid Papillary Carcinoma - Follicular (>= 99% follicular patterned)	Thyroid Gland Follicular Carcinoma	C8054	102
THCA	Thyroid Papillary Carcinoma - Tall Cell (>= 50% tall cell features)	Thyroid Gland Papillary Carcinoma	C4035	37
THYM	Thymoma; Type A	Thymoma	C3411	15
THYM	Thymoma; Type AB	Thymoma	C3411	38
THYM	Thymoma; Type A Thymoma; Type AB	Thymoma	C3411	2
THYM	Thymoma; Type B1	Thymoma	C3411	14
THYM	Thymoma; Type B1 Thymoma; Type B2	Thymoma	C3411	1

THYM	Thymoma; Type B2	Thymoma	C3411	28
THYM	Thymoma; Type B2 Thymoma; Type B3	Thymoma	C3411	3
THYM	Thymoma; Type B3	Thymoma	C3411	12
THYM	Thymoma; Type C	Thymoma	C3411	11
UCEC	Endometrioid endometrial adenocarcinoma	Endometrial Endometrioid Adenocarcinoma	C6287	411
UCEC	Mixed serous and endometrioid	Endometrial Mixed Cell Adenocarcinoma	C40153	22
UCEC	Serous endometrial adenocarcinoma	Endometrial Serous Adenocarcinoma	C27838	115
UCS	Uterine Carcinosarcoma/ MMMT: Heterologous Type	Uterine Carcinosarcoma	C42700	20
UCS	Uterine Carcinosarcoma/ Malignant Mixed Mullerian Tumor (MMMT): NOS	Uterine Carcinosarcoma	C42700	24
UCS	Uterine Carcinosarcoma/MMMT: Homologous Type	Uterine Carcinosarcoma	C42700	13
UVM	Epithelioid Cell	Uveal Melanoma	C7712	13
UVM	Epithelioid Cell Spindle Cell	Uveal Melanoma	C7712	21
UVM	Spindle Cell	Uveal Melanoma	C7712	30
UVM	Spindle Cell Epithelioid Cell	Uveal Melanoma	C7712	16

Appendix 2: Patient demographics of included tumor types

	n	age (median [IQR])	gender = F/M (%)	race (%)				death n (%)	progression n (%)	TMB count (median [IQR])	NonSyn count (median [IQR])	SNP count (median [IQR])	INDEL count (median [IQR])
				WHITE	BLACK	ASIAN	OTHERS						
Adenocarcinoma, Pancreas	173	65.00 [57.00, 73.00]	77/96 (44.5/55.5)	151 (87.3)	7 (4.0)	10 (5.8)	5 (2.9)	95 (54.9)	108 (62.4)	53.00 [40.00, 71.00]	35.00 [25.00, 46.00]	32.00 [23.00, 43.00]	2.00 [1.00, 3.00]
Astrocytoma	195	39.00 [32.00, 51.00]	86/109 (44.1/55.9)	181 (92.8)	8 (4.1)	2 (1.0)	4 (2.1)	58 (29.7)	90 (46.2)	43.00 [31.00, 58.00]	30.00 [22.00, 39.00]	27.00 [20.00, 36.00]	2.00 [1.00, 3.00]
Cervical Squamous Cell Carcinoma	240	47.00 [39.00, 57.00]	240/0 (100.0/0.0)	158 (65.8)	27 (11.2)	14 (5.8)	41 (17.1)	57 (23.8)	54 (22.5)	166.00 [101.00, 290.75]	92.00 [60.00, 156.00]	86.50 [55.00, 149.50]	2.00 [1.00, 4.00]
Clear Cell Renal Cell Carcinoma	368	60.00 [51.00, 69.00]	136/232 (37.0/63.0)	301 (81.8)	54 (14.7)	7 (1.9)	6 (1.6)	98 (26.6)	96 (26.1)	76.00 [56.00, 99.00]	52.00 [37.00, 66.00]	45.00 [33.00, 58.00]	6.00 [3.00, 9.00]
Colorectal Adenocarcinoma	545	67.00 [57.00, 76.00]	261/284 (47.9/52.1)	279 (51.2)	62 (11.4)	12 (2.2)	192 (35.2)	115 (21.1)	143 (26.2)	167.00 [125.00, 239.00]	106.00 [80.00, 156.00]	98.00 [74.00, 148.00]	5.00 [3.00, 8.00]
Cutaneous Melanoma	453	58.00 [48.00, 71.00]	170/283 (37.5/62.5)	431 (95.1)	0 (0.0)	12 (2.6)	10 (2.2)	214 (47.2)	307 (67.8)	692.00 [312.00, 1469.00]	415.00 [193.00, 906.00]	401.00 [186.00, 872.00]	3.00 [1.00, 5.00]
Endometrial Endometrioid Adenocarcinoma	397	62.00 [55.00, 70.00]	397/0 (100.0/0.0)	283 (71.3)	66 (16.6)	17 (4.3)	31 (7.8)	48 (12.1)	73 (18.4)	246.00 [79.00, 1216.00]	144.00 [50.00, 652.00]	117.00 [45.00, 500.00]	9.00 [4.00, 93.00]
Endometrial Serous Adenocarcinoma	111	68.00 [63.00, 73.00]	111/0 (100.0/0.0)	65 (58.6)	32 (28.8)	3 (2.7)	11 (9.9)	33 (29.7)	40 (36.0)	85.00 [65.00, 155.00]	54.00 [41.50, 94.00]	49.00 [38.50, 86.00]	3.00 [2.00, 4.00]
Gastric Adenocarcinoma	431	67.00 [58.00, 73.00]	157/274 (36.4/63.6)	272 (63.1)	13 (3.0)	88 (20.4)	58 (13.5)	168 (39.0)	139 (32.3)	171.00 [106.00, 366.00]	117.00 [72.00, 248.00]	108.00 [65.50, 235.50]	5.00 [2.00, 12.00]
Glioblastoma	400	61.00	147/253	342	42	6	10	308	319	77.00	51.00	47.00	2.00

		[52.00, 69.00]	(36.8/63.2)	(85.5)	(10.5)	(1.5)	(2.5)	(77.0)	(79.8)	[62.00, 98.25]	[40.00, 67.00]	[38.00, 62.00]	[1.00, 4.00]
Head and Neck Squamous Cell Carcinoma	509	61.00 [53.00, 69.00]	139/370 (27.3/72.7)	436 (85.7)	47 (9.2)	10 (2.0)	16 (3.1)	220 (43.2)	196 (38.5)	163.00 [100.00, 271.00]	108.00 [65.00, 178.00]	100.00 [60.00, 168.00]	4.00 [2.00, 8.00]
Hepatocellular Carcinoma	364	61.00 [51.00, 69.00]	119/245 (32.7/67.3)	181 (49.7)	16 (4.4)	157 (43.1)	10 (2.7)	123 (33.8)	179 (49.2)	137.50 [97.00, 186.00]	83.00 [61.00, 117.00]	77.00 [57.00, 107.00]	5.00 [3.00, 7.00]
Infiltrating Ductal Breast Carcinoma	759	57.00 [48.00, 66.00]	749/10 (98.7/1.3)	496 (65.3)	140 (18.4)	46 (6.1)	77 (10.1)	101 (13.3)	95 (12.5)	68.00 [44.00, 115.50]	43.00 [28.00, 74.00]	40.00 [25.00, 69.00]	3.00 [1.00, 4.00]
Invasive Lobular Breast Carcinoma	167	62.00 [53.00, 69.00]	167/0 (100.0/0.0)	136 (81.4)	9 (5.4)	7 (4.2)	15 (9.0)	17 (10.2)	16 (9.6)	57.00 [38.00, 99.00]	33.00 [24.00, 59.50]	31.00 [21.00, 54.00]	2.00 [1.00, 4.00]
Lung Adenocarcinoma	508	66.00 [59.00, 72.75]	270/238 (53.1/46.9)	389 (76.6)	53 (10.4)	8 (1.6)	58 (11.4)	182 (35.8)	208 (40.9)	300.00 [120.75, 605.25]	202.50 [83.00, 407.50]	192.00 [77.75, 381.00]	7.00 [3.00, 14.00]
Lung Squamous Cell Carcinoma	479	68.00 [62.00, 73.00]	127/352 (26.5/73.5)	339 (70.8)	28 (5.8)	9 (1.9)	103 (21.5)	208 (43.4)	140 (29.2)	341.00 [240.00, 501.50]	230.00 [159.50, 338.00]	217.00 [152.50, 316.00]	8.00 [5.00, 13.00]
Muscle-Invasive Bladder Carcinoma	407	69.00 [60.00, 76.00]	107/300 (26.3/73.7)	323 (79.4)	23 (5.7)	43 (10.6)	18 (4.4)	179 (44.0)	174 (42.8)	265.00 [139.00, 463.50]	173.00 [91.00, 301.50]	166.00 [84.50, 290.00]	4.00 [3.00, 7.00]
Oligoastrocytoma	195	45.00 [34.00, 54.00]	89/106 (45.6/54.4)	177 (90.8)	7 (3.6)	5 (2.6)	6 (3.1)	46 (23.6)	73 (37.4)	39.00 [28.00, 51.50]	27.00 [18.00, 35.00]	24.00 [16.00, 31.00]	2.00 [1.00, 3.00]
Oligodendroglioma	134	38.00 [30.00, 49.75]	60/74 (44.8/55.2)	125 (93.3)	7 (5.2)	1 (0.7)	1 (0.7)	27 (20.1)	41 (30.6)	35.00 [26.00, 50.00]	25.00 [19.00, 35.75]	21.50 [17.00, 33.00]	2.00 [1.00, 3.00]
Ovarian Serous Adenocarcinoma	408	59.00 [51.00, 68.00]	408/0 (100.0/0.0)	347 (85.0)	26 (6.4)	14 (3.4)	21 (5.1)	233 (57.1)	276 (67.6)	100.50 [71.00, 146.00]	67.00 [47.00, 99.00]	61.50 [42.75, 90.25]	4.00 [3.00, 7.00]
Papillary Renal Cell Carcinoma	280	61.00 [53.25, 71.00]	77/203 (27.5/72.5)	198 (70.7)	60 (21.4)	6 (2.1)	16 (5.7)	40 (14.3)	54 (19.3)	101.00 [65.75, 136.00]	66.00 [41.75, 88.00]	55.00 [36.00, 73.25]	8.00 [4.00, 12.00]

Pheochromocytoma	151	46.00 [35.00, 58.50]	85/66 (56.3/43.7)	127 (84.1)	14 (9.3)	5 (3.3)	5 (3.3)	4 (2.6)	15 (9.9)	14.00 [9.50, 19.00]	9.00 [6.00, 13.00]	8.00 [6.00, 12.00]	0.00 [0.00, 1.00]
Prostate Acinar Adenocarcinoma	483	61.00 [56.00, 66.00]	0/483 (0.0/100.0)	146 (30.2)	7 (1.4)	2 (0.4)	328 (67.9)	9 (1.9)	89 (18.4)	39.00 [29.50, 52.00]	26.00 [20.00, 35.00]	24.00 [18.00, 32.00]	1.00 [0.00, 3.00]
Thymoma	122	60.50 [50.00, 68.75]	58/64 (47.5/52.5)	101 (82.8)	6 (4.9)	13 (10.7)	2 (1.6)	9 (7.4)	22 (18.0)	28.00 [19.00, 41.00]	15.00 [10.00, 23.75]	13.00 [9.00, 22.00]	1.00 [0.00, 1.00]
Thyroid Gland Papillary Carcinoma	391	46.00 [34.00, 58.00]	287/104 (73.4/26.6)	275 (70.3)	20 (5.1)	46 (11.8)	50 (12.8)	13 (3.3)	44 (11.3)	15.00 [10.00, 23.00]	10.00 [7.00, 16.00]	9.00 [6.00, 15.00]	0.00 [0.00, 1.00]

Appendix 3: Patient demographics of excluded tumor types

	n	age (median IQR])	gender = F/ M (%)	race (%)				death n (%)	progression n (%)	TMB count (median [IQR])	NonSyn count (median [IQR])	SNP count (median [IQR])	INDEL count (median [IQR])
				WHITE	BLACK	ASIAN	OTHERS						
Adrenal Cortex Carcinoma	92	48.50 [35.50, 60.00]	60/32 (65.2/34.8)	78 (84.8)	1 (1.1)	2 (2.2)	11 (12.0)	34 (37.0)	49 (53.3)	39.50 [26.25, 87.25]	26.00 [17.75, 54.00]	24.00 [15.75, 47.25]	2.00 [1.00, 3.00]
Breast Carcinoma	98	57.00 [48.25, 70.00]	97/1 (99.0/1.0)	78 (79.6)	13 (13.3)	6 (6.1)	1 (1.0)	20 (20.4)	20 (20.4)	50.00 [33.00, 84.50]	30.00 [21.25, 54.00]	26.00 [19.00, 48.75]	2.00 [1.00, 3.00]
Cervical Adenocarcinoma	44	44.00 [36.75, 55.25]	44/0 (100.0/0.0)	35 (79.5)	1 (2.3)	5 (11.4)	3 (6.8)	10 (22.7)	13 (29.5)	122.00 [78.75, 257.75]	72.50 [45.50, 137.50]	69.00 [41.75, 120.75]	2.00 [1.00, 3.25]
Cervical Adenosquamous Carcinoma	7	34.00 [33.50, 45.00]	7/0 (100.0/0.0)	3 (42.9)	2 (28.6)	0 (0.0)	2 (28.6)	1 (14.3)	1 (14.3)	83.00 [68.50, 118.00]	40.00 [38.00, 73.00]	38.00 [34.00, 66.50]	2.00 [2.00, 4.00]
Cholangiocarcinoma	36	66.50 [56.50, 72.00]	20/16 (55.6/44.4)	31 (86.1)	2 (5.6)	3 (8.3)	0 (0.0)	18 (50.0)	20 (55.6)	77.00 [58.50, 130.50]	41.00 [31.75, 68.75]	38.50 [28.00, 66.25]	3.00 [2.00, 4.25]
Chromophobe Renal Cell Carcinoma	65	50.00 [42.00, 61.00]	27/38 (41.5/58.5)	57 (87.7)	4 (6.2)	2 (3.1)	2 (3.1)	9 (13.8)	11 (16.9)	31.00 [26.00, 38.00]	19.00 [16.00, 26.00]	18.00 [15.00, 24.00]	1.00 [0.00, 2.00]
Desmoid-Type Fibromatosis	2	40.00 [32.00, 48.00]	1/1 (50.0/50.0)	2 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)	0 (0.0)	1 (50.0)	32.50 [30.25, 34.75]	17.50 [14.75, 20.25]	17.00 [14.50, 19.50]	0.00 [0.00, 0.00]
Diffuse Large B-Cell Lymphoma	37	57.00 [46.00, 67.00]	21/16 (56.8/43.2)	20 (54.1)	0 (0.0)	17 (45.9)	0 (0.0)	5 (13.5)	8 (21.6)	176.00 [120.00, 273.00]	109.00 [82.00, 160.00]	104.00 [79.00, 148.00]	4.00 [3.00, 7.00]
Endometrial Mixed Cell Adenocarcinoma	22	67.00 [60.00, 73.50]	22/0 (100.0/0.0)	13 (59.1)	7 (31.8)	0 (0.0)	2 (9.1)	6 (27.3)	8 (36.4)	91.50 [63.00, 1475.75]	59.00 [37.75, 756.25]	54.00 [34.50, 527.25]	4.00 [2.25, 55.00]
Esophageal	88	69.50	12/76	71	0	1	16	44	44 (50.0)	220.50	125.00	112.50	5.00

Adenocarcinoma		[58.00, 77.00]	(13.6/86.4)	(80.7)	(0.0)	(1.1)	(18.2)	(50.0)		[175.50, 329.00]	[98.50, 191.00]	[89.00, 170.00]	[3.00, 8.00]
Esophageal Squamous Cell Carcinoma	97	57.00 [51.00, 64.00]	15/82 (15.5/84.5)	43 (44.3)	6 (6.2)	45 (46.4)	3 (3.1)	33 (34.0)	43 (44.3)	164.00 [129.00, 241.00]	94.00 [69.00, 127.00]	84.00 [62.00, 117.00]	4.00 [3.00, 6.00]
Leiomyosarcoma	94	58.00 [52.25, 66.75]	62/32 (66.0/34.0)	79 (84.0)	11 (11.7)	2 (2.1)	2 (2.1)	37 (39.4)	56 (59.6)	73.50 [54.00, 96.75]	40.50 [29.00, 53.75]	37.00 [26.25, 47.75]	2.00 [1.00, 3.00]
Liposarcoma	48	64.00 [53.00, 76.25]	16/32 (33.3/66.7)	46 (95.8)	0 (0.0)	1 (2.1)	1 (2.1)	19 (39.6)	24 (50.0)	63.50 [47.25, 86.25]	37.00 [26.00, 48.00]	33.00 [25.00, 44.50]	1.00 [0.00, 2.00]
Malignant Peripheral Nerve Sheath Tumor	10	42.00 [29.50, 52.00]	5/5 (50.0/50.0)	9 (90.0)	1 (10.0)	0 (0.0)	0 (0.0)	4 (40.0)	7 (70.0)	70.50 [54.00, 87.00]	39.00 [31.25, 48.00]	35.00 [28.75, 46.25]	1.50 [1.00, 2.00]
Mesothelioma	79	64.00 [57.00, 69.00]	15/64 (19.0/81.0)	77 (97.5)	1 (1.3)	1 (1.3)	0 (0.0)	67 (84.8)	55 (69.6)	45.00 [32.50, 56.00]	28.00 [19.50, 35.50]	24.00 [17.00, 32.50]	2.00 [1.00, 3.00]
Myxofibrosarcoma	25	60.00 [54.00, 75.00]	14/11 (56.0/44.0)	20 (80.0)	2 (8.0)	1 (4.0)	2 (8.0)	9 (36.0)	12 (48.0)	112.00 [72.00, 143.00]	59.00 [36.00, 78.00]	54.00 [33.00, 70.00]	2.00 [0.00, 3.00]
Pancreatic Carcinoma	4	59.50 [55.50, 66.75]	2/2 (50.0/50.0)	3 (75.0)	0 (0.0)	1 (25.0)	0 (0.0)	2 (50.0)	0 (0.0)	87.00 [43.75, 119.50]	58.50 [25.00, 86.50]	54.00 [21.25, 83.00]	1.00 [0.00, 2.50]
Paraganglioma	33	46.00 [39.00, 57.00]	17/16 (51.5/48.5)	26 (78.8)	6 (18.2)	1 (3.0)	0 (0.0)	4 (12.1)	8 (24.2)	13.00 [9.00, 21.00]	9.00 [7.00, 14.00]	8.00 [6.00, 14.00]	0.00 [0.00, 1.00]
Prostate Adenocarcinoma	15	61.00 [59.00, 66.50]	0/15 (0.0/100.0)	0 (0.0)	0 (0.0)	0 (0.0)	15 (100.0)	1 (6.7)	4 (26.7)	48.00 [38.00, 58.50]	32.00 [28.00, 40.00]	30.00 [27.00, 37.00]	2.00 [0.50, 2.50]
Synovial Sarcoma	10	29.50 [27.25, 38.75]	6/4 (60.0/40.0)	10 (100.0)	0 (0.0)	0 (0.0)	0 (0.0)	4 (40.0)	3 (30.0)	46.00 [37.50, 51.75]	25.00 [17.25, 32.50]	23.00 [16.25, 30.25]	1.00 [0.25, 1.00]
Testicular Non-Seminomatous Germ Cell Tumor	64	28.00 [23.00, 33.25]	0/64 (0.0/100.0)	58 (90.6)	3 (4.7)	1 (1.6)	2 (3.1)	3 (4.7)	25 (39.1)	25.00 [16.00, 30.25]	15.50 [10.00, 20.25]	14.00 [8.75, 17.25]	1.00 [1.00, 2.00]

Testicular Seminoma	70	32.50 [28.00, 38.00]	0/70 (0.0/100.0)	61 (87.1)	3 (4.3)	3 (4.3)	3 (4.3)	1 (1.4)	14 (20.0)	20.00 [15.00, 26.00]	12.50 [9.00, 17.00]	11.50 [8.00, 14.75]	1.00 [0.00, 2.00]
Thyroid Gland Follicular Carcinoma	100	47.00 [37.75, 58.00]	76/24 (76.0/24.0)	45 (45.0)	6 (6.0)	4 (4.0)	45 (45.0)	1 (1.0)	6 (6.0)	15.50 [12.00, 21.25]	11.00 [8.75, 15.00]	10.00 [7.75, 13.25]	0.00 [0.00, 1.00]
Undifferentiated Pleomorphic Sarcoma	50	69.00 [59.25, 78.00]	26/24 (52.0/48.0)	45 (90.0)	3 (6.0)	2 (4.0)	0 (0.0)	15 (30.0)	24 (48.0)	88.00 [70.75, 118.00]	47.00 [38.00, 68.00]	45.00 [36.00, 63.50]	1.00 [0.00, 3.00]
Uterine Carcinosarcoma	57	68.00 [62.00, 76.00]	57/0 (100.0/0.0)	44 (77.2)	9 (15.8)	3 (5.3)	1 (1.8)	35 (61.4)	37 (64.9)	66.00 [58.00, 84.00]	46.00 [40.00, 57.00]	43.00 [36.00, 52.00]	3.00 [2.00, 4.00]
Uveal Melanoma	79	60.00 [51.00, 74.50]	35/44 (44.3/55.7)	55 (69.6)	0 (0.0)	0 (0.0)	24 (30.4)	22 (27.8)	29 (36.7)	16.00 [13.00, 21.00]	11.00 [8.50, 15.00]	11.00 [7.00, 14.00]	1.00 [0.00, 1.00]

Curriculum Vita

JULIA KUNG

EDUCATION

Johns Hopkins University (Baltimore, MD) **2018-2020**

Master of Science: Health Sciences Informatics (2-year research program)

- **Academics:** 3.99/4 GPA
- **Relevant courses:** Applied Clinical Informatics, Medical Students Informatics Education, Computational Genomics Methods, Database Querying in Health, Introduction to Data Management
- **Teaching Assistant Positions:** Clinical Data Analysis with Python, Introduction to Health Informatics, Seminar & Grand Rounds

Taipei Medical University (Taipei, Taiwan) **2013-2017**

Bachelor of Science: Medical Laboratory Science and Biotechnology

- **Academics:** 3.95/4 GPA
- **Graduation President Award:** top 1% of graduating class
- **Relevant courses:** Medical Genetics, Cell Biology, Molecular Biology, Clinical Serology & Immunology, Microbiology, Pathology, Clinical Data Interpretation

ACADEMIC RESEARCHES

Graduate student researcher **2018 Aug -Present**

Dr. Alexander Baras Lab, Johns Hopkins University

Analyzed survival probability of pan-cancer (breast, lung, colon, prostate, and melanoma cancer) patients using programming language R and computed aggregate genomic metrics to improve precision of cancer prognosis.

Graduate student researcher **2019 Mar – 2019 Oct**

Dr. Marion Ball and Dr. George Kim, Johns Hopkins University

Collaborate with IBM Watson Health Team and evaluated the complementarity of EHR and Claims in assessing diagnosis from R-health Direct Primary Care patient data using programming language R and SQL.

SKILLS AND LICENSE

Computer skills: R, Python, SQL, and Microsoft Excel.
License: Medical Technologist
Laboratory skills: Phlebotomy, Western Blot, cell/bacteria/virus culture, PCR, ELISA
Other Skills: regulatory writing, IRB review process

CLINICAL EXPERIENCE

Children's Healthcare of Atlanta, Observation Program **2017 Apr – 2017 May**
Department of Pathology and Laboratory Services

Tri-Service General Hospital, Intern **2016 Aug – 2016 Dec**
Rotated in Clinical Biochemistry Lab, Clinical Microbiology Lab, Clinical Hematology Lab, Blood Bank, Clinical Microscopic Examination Lab, Clinical Serology & Immunology Lab, Clinical Physiology Lab, Pathology Lab, and Molecular Diagnostic Lab
Practicum in phlebotomy – blood and urine routine test

COURSE PROJECTS

Health Sciences Informatics, Knowledge Engineering and Decision Support, JHU
Planned a decision support system than embodies the HIMSS framework
2020 Jan - 2020 Apr
Designed a decision support system for rare blood typing result interpretation, including Bombay/ Para-Bombay testing and discrepancy-investigating workflow, user interface, and evaluation.

HIT Standards and Systems Interoperability, JHU
Developed Functional Requirement Analysis Document (FRAD)
2018 Aug - 2018 Oct
Composed a functional requirement analysis document for developing an early warning and surveillance system for gastrointestinal disease.

LEADERSHIP AND OUTREACH EXPERIENCE

Taiwanese Student Association – JHU SOM Representative **2019 May – 2020 May**

Club leader of Taipei Medical University Art Club **2014 Aug – 2015 Jun**
Awarded Excellent Performance in Club Evaluation

Immigrant Children Volunteer Tutor **2013 Aug – 2014 Jun**
Awarded Best Volunteer Honor